Applying phylogenomics to understand the emergence of Shiga Toxin producing Escherichia coli O157:H7 strains causing severe human disease in the United Kingdom.

Microbial Genomics

Applying phylogenomics to understand the emergence of Shiga Toxin producing Escherichia coli O157:H7 strains causing severe human disease in the United Kingdom.

--Manuscript Draft--

Manuscript Number:	MGEN-D-15-00009R2
Full Title:	Applying phylogenomics to understand the emergence of Shiga Toxin producing Escherichia coli O157:H7 strains causing severe human disease in the United Kingdom.
Article Type:	Research Paper
Section/Category:	Microbial evolution and epidemiology: Population Genomics
Corresponding Author:	Tim Dallman Public Health England Colindale UNITED KINGDOM
First Author:	Tim Dallman
Order of Authors:	Tim Dallman
	Philip Ashton
	Lisa Byrne
	Neil Perry
	Liljana Petrovska
	Richard Ellis
	Lesley Allison
	Mary Hanson
	Anne Holmes
	George Gunn
	Margo Chase-Topping
	Mark Woolhouse
	Kathie Grant
	David Gally
	John Wain
	Claire Jenkins
Abstract:	Shiga Toxin producing Escherichia coli (STEC) O157:H7 is a recently emerged zoonotic pathogen with considerable morbidity. Since the serotype emerged in the 1980s, research has focussed on unravelling the evolutionary events from the E. coli O55:H7 ancestor to the contemporaneous globally dispersed strains. In this study the genomes of over 1000 isolates from human clinical cases and cattle, spanning the history of STEC O157:H7 in the United Kingdom were sequenced. Phylogenetic analysis reveals the ancestry, key acquisition events and global context of the strains. Dated phylogenies estimate the time to the most recent common ancestor of the current circulating global clone to 175 years ago, followed by rapid diversification. We show the acquisition of specific virulence determinates occurred relatively recently and coincides with its recent detection in the human population. Using clinical outcome data from 493 cases of STEC O157:H7 we assess the relative risk of severe disease including HUS from each of the defined clades in the population and show the dramatic effect Shiga toxin complement has on virulence. We describe two strain replacement events that have occurred in the cattle population in the UK over the last 30 years; one resulting in a highly virulent strain that has accounted for the majority of clinical cases

	in the UK over the last decade. This work highlights the need to understand the selection pressures maintaining Shiga-toxin encoding bacteriophages in the ruminant reservoir and the study affirms the requirement for close surveillance of this pathogen in both ruminant and human populations.

MICROBIAL GENOMICS

Research paper template

Applying phylogenomics to understand the emergence of Shiga Toxin producing Escherichia coli O157:H7 strains causing severe human disease in the United Kingdom.

Timothy J. Dallman[1*], Philip M. Ashton[1], Lisa Byrne[1], Neil T. Perry[1], Liljana Petrovska[3], Richard Ellis[3], Lesley Allison[5], Mary Hanson[5], Anne Holmes[5], George J. Gunn[7], Margo E. Chase-Topping[6], Mark E. J. Woolhouse[6], Kathie A. Grant[1], David L. Gally[4], John Wain[2*], Claire Jenkins[1].

[1]Public Health England, 61 Colindale Avenue, London, NW9 5EQ

[2]University of East Anglia, Norwich, NR4 7TJ

[3]Animal Laboratories and Plant Health Agency, Woodham Lane, Surrey, KT15 3NB

[4]Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, UK, EH25 9RG.

[5]Scottish *E. coli* O157/VTEC Reference Laboratory, Department of Laboratory Medicine, Royal Infirmary of Edinburgh, 51 Little France Crescent, Edinburgh EH16 4SA.

[6]Centre for Immunity, Infection and Evolution, Kings Buildings, University of Edinburgh, Edinburgh, UK, EH9 3FL.

[7] Future Farming Systems, R&D Division, SRUC, Drummondhill, Stratherrick Rd., Inverness, Scotland, UK, IV2 4JZ

***Corresponding author – tim.dallman@phe.gov.uk**

ABSTRACT

Shiga Toxin producing *Escherichia coli* (STEC) O157:H7 is a recently emerged zoonotic pathogen with considerable morbidity. Since the emergence of this serotype in the 1980s, research has focussed on unravelling the evolutionary events from the *E. coli* O55:H7 ancestor to the contemporaneous globally dispersed strains observed today. In this study the genomes of over one thousand isolates from both human clinical cases and cattle, spanning the history of STEC O157:H7 in the United Kingdom were sequenced. Phylogenetic analysis reveals the ancestry, key acquisition events and global context of the strains. Dated phylogenies estimate the time to evolution of the most recent common ancestor of the current circulating global clone to be 175 years ago. This event was followed by rapid diversification. We show the acquisition of specific virulence determinates has

occurred relatively recently and coincides with its recent detection in the human population. We used clinical outcome data from 493 cases of STEC O157:H7 to assess the relative risk of severe disease including HUS from each of the defined clades in the population and show the dramatic effect Shiga toxin repertoire has on virulence. We describe two strain replacement events that have occurred in the cattle population in the United Kingdom over the last 30 years; one resulting in a highly virulent strain that has accounted for the majority of clinical cases in the United Kingdom over the last decade. There is a need to understand the selection pressures maintaining Shiga-toxin encoding bacteriophages in the ruminant reservoir and the study affirms the requirement for close surveillance of this pathogen in both ruminant and human populations.

DATA SUMMARY

FASTQ sequences were deposited in the NCBI Short Read Archive under the BioProject PRJNA248042 (http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA248042)

Supplementary Table 5 is available at the following git repository

https://github.com/timdallman/phylogenomics_stec.git

I/We confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. ☑

IMPACT STATEMENT

In this article we analyse over 1000 Shiga Toxin producing *Escherichia coli* (STEC) O157:H7 genomes from animal and clinical isolates collected over the past three decades and present for the first time a comprehensive population structure of STEC O157:H7. Using phylogenetic methods we have examined the origin and dispersal of this zoonotic pathogen and show how historical worldwide dissemination followed by regional expansion in native cattle populations gives rise to the extant diversity seen today. By comparing clinical outcome data of nearly 500 human cases we comprehensively assess the association between phylogenetic grouping, acquisition and loss of specific subtypes of Shiga toxin and severe disease. With this analysis we show specific circulating strains have >5 fold increase risk of severe disease than the ancestral STEC O157:H7 genotype. Finally we show that recent strain replacement has occurred in Great Britain shaping the diversity of STEC O157:H7 observed today and introducing a high virulence clone into the British cattle population.

INTRODUCTION

Shiga Toxin producing Escherichia coli (STEC) O157:H7 is a globally dispersed pathogen that, whilst generally asymptomatic in its ruminant host, can cause severe outbreaks of gastroenteritis, haemorrhagic colitis and haemolytic uraemic syndrome in humans (Akashi et al., 1994; Centers for Disease Control and Prevention (CDC), 2006; Ihekweazu et al., 2012). Contemporary STEC O157:H7 represent a monomorphic clone(Whittam et al., 1988) characterised by particular phenotypic properties including the inability to ferment sorbitol and produce β-glucuronidase. Over the course of its evolution, STEC O157:H7 has acquired several virulence determinants including two types of Shiga toxins (Stx1 and Stx2) encoded on lambdoid bacteriophages (Scotland et al., 1985), a myriad of effector proteins(Lai et al., 2013; Tobe et al., 2006) and a virulence plasmid containing genes for a type II secretion system and a haemolysin (Schmidt et al., 1994). It is postulated that the current clone arose with the transfer of the O157 *rfb* and *gnd* genes that specify the structure of lipopolysaccharide side chains that comprise the somatic (O) antigens into a *stx2* containing *E. coli* O55:H7 strain that had an enhanced capacity for host colonisation mediated by the locus of enterocyte effacement (LEE) pathogenicity island (Wick et al., 2005). A step-wise sequence of events involving the loss of the ability to utilise sorbitol, lysogenisation by an *stx1* containing phage and inactivation of the gene encoding the β-glucuronidase *uidA* is hypothesised to have given rise to the currently circulating clone (Feng et al., 1998), with distinct subpopulations formed by less common non-motile O157:H- strains and strains that retained the ability to express β-glucuronidase.

Despite high levels of relatedness of the non-sorbitol fermenting, β-glucuronidase negative STEC O157:H7 strains, it has long been realised that distinct lineages exist within the population. It is suggested that these arose from the result of geographic spread of an ancestral clone and subsequent regional expansion (Kim et al., 2001; Yang et al., 2004). Identified subpopulations have also been found to be unequally distributed in the cattle and human populations with lineage I being more prevalent among human clinical isolates and lineage II more associated with the animal host (Yang et al., 2004). Subsequent studies revealed differences between the two lineages including Stx-encoding bacteriophage (StxΦ) insertion sites (Besser et al., 2007), *stx2* expression (Dowd and Williams, 2008), stress resistance (Lee et al., 2012), as well as lineage specific polymorphisms (Bono et al., 2007). Further characterisation of genomic differences between these two lineages identified an intermediate genogroup termed lineage I/II (Zhang et al., 2007). To investigate the propensity of different STEC O157:H7 strains to cause serious illness, further sub-typing schemes have been developed which sub-divided the population into 9 clades based on single nucleotide polymorphisms (Manning et al., 2008; Riordan et al., 2008) with clade 8 associated with two large outbreaks of Haemolytic Uremic Syndrome (HUS) (Manning et al., 2008). Subsequent *in vitro* studies showed varied adherence and virulence factor expression between different clades (Abu-Ali et al., 2010) and whole genome studies elucidated further potential virulence determinants (Eppinger et al., 2011a). The use of clade genotyping provided further evidence that the diversity within STEC O157:H7 is globally distributed (Mellor et al., 2013; Yokoyama et al., 2012).

Several groups have used the clade description of the STEC O157:H7 population to further speculate on the evolutionary path that has given rise to the current diversity(Kyle et al., 2012; Leopold et al., 2009; Yokoyama et al., 2012). The current model suggests that β-glucuronidase positive, non-sorbitol fermenting STEC O157:H7 (clade 9) are ancestral to lineage II and the intermediate lineage

I/II (which overlap with clades 8-5) which themselves are ancestral to lineage I (clades 5-1). The nature of the paraphyletic evolution of these lineages however remains unknown.

The United Kingdom (UK) has a comparatively high human infection rate with STEC O157(Chase-Topping et al., 2008) and this has remained relatively constant over the last decade. In the UK, STEC O157 strains are subtyped by determining sensitivity to a specific panel of 16 typing phages, a phage typing scheme developed in Canada and adopted by several European countries(Ahmed et al., 1987; Khakhria et al., 1990). Over the last decade in England, Scotland and Wales, phage type (PT) 21/28 strains have been most commonly associated with severe human infection and more recent research has indicated that these strains are more likely to be associated with high excretion levels from cattle; known as supershedding(Chase-Topping et al., 2008). Previously, the most common phage type in England, Scotland and Wales was PT2 until it decreased year after year from 1998 (see Figure 1). The nature of this strain replacement and how PT21/28, PT2 and other common phage types, such as PT8 and PT32 are associated with each other and to the lineages defined above was not understood. In this study we present the population structure of STEC O157:H7 from a UK perspective using genome sequencing of over 1000 animal and clinical isolates collected over the past three decades. Using phylogenetic methods we have examined the origin and dispersal of this zoonotic pathogen and estimated approximate evolutionary timescales that have led to the emergence of an expanded virulent cluster that accounts for a significant proportion of the human STEC disease in the UK.

METHODS

Strain Selection

1075 strains of STEC O157 from clinical and animal isolates from England, Northern Ireland, Wales & Scotland collected from 1985 to 2014 were selected for sequencing. These represented 25 phage types. Ninety-five cattle strains were STEC O157:H7 isolates selected for sequencing from Scottish cattle strains collected as part of 'The Wellcome Foundation International Partnership Research Award in Veterinary Epidemiology' (IPRAVE) study on the basis of regional and genotypic diversity. 54 sequences were downloaded from public repositories including the oldest sequenced STEC O157(Sanjar et al., 2014).

Genome Sequencing and Sequence Analysis

Genomic DNA was fragmented and tagged for multiplexing with Nextera XT DNA Sample Preparation Kits (Illumina) and sequenced at the Animal Laboratories and Plant Health Agency using the Illumina GAII platform with 2x150bp reads. Short reads were quality trimmed(Bolger et al., 2014) and mapped to the reference STEC O157 strain *Sakai* (Genbank accession BA000007) using BWA-SW(Li and Durbin, 2010). The Sequence Alignment Map output from BWA was sorted and indexed to produce a Binary Alignment Map (BAM) using Samtools(Li et al., 2009). GATK2(McKenna et al., 2010) was used to create a Variant Call Format (VCF) file from each of the BAMs, which were further parsed to extract only single nucleotide polymorphism (SNP) positions which were of high quality (MQ>30, DP>10, GQ>30, Variant Ratio >0.9). Pseudosequences of polymorphic positions were used to create maximum likelihood trees using RaxML(Stamatakis, 2014). Pair-wise SNP distances between each pseudosequence were calculated. Spades version 2.5.1(Bankevich et al., 2012) was run using careful mode with kmer sizes 21, 33, 55 and 77 to produce *de novo* assemblies of the sequenced paired-end fastq files. FASTQ sequences were deposited in the NCBI Short Read Archive under the BioProject PRJNA248042.

SNP Clustering

Hierarchical single linkage clustering was performed on the pairwise SNP difference between all strains at various distance thresholds (Δ250, Δ100, Δ50, Δ25, Δ10, Δ5, Δ0). The result of the clustering is a SNP address that can be used to describe the population structure based on clonal groups.

Recombination

Recombination analysis was performed using BRATNEXTGEN(Marttinen et al., 2012). Representatives from Δ50 SNP clusters were randomly selected and whole genome alignment produced relative to the reference strain *Sakai*. From the proportion of shared ancestry generated by BRATNEXTGEN the dataset was partitioned into 18 clusters. Recombination between and within these clusters was calculated over 20 iterations and the significance estimated over 100 replicates. Detected recombinant segments were deemed significant with a p-value < 0.05.

Timed phylogenies

Timed phylogenies were constructed using BEAST-MCMC. v1.80(Drummond et al., 2012) and after first confirming a temporal signal using Path-O-Gen(Drummond et al., 2012). Alternative clock models and population priors were computed and their suitability assessed based on Bayes Factor (BF) tests. The highest supported model was a relaxed lognormal clock rate under a constant population size. All models were run with a chain length of 1 billion. A maximum clade credibility tree was constructed using TreeAnnotator v1.75.

195

196 Shiga toxin subtyping

197

198 Shiga toxin subtyping was performed as described by Ashton and colleagues (Ashton et al., 2015).

199

200 Stx-associated bacteriophage insertion (SBI)

201

202 The integration of shiga toxin carrying prophage into the host genome has been characterised into
203 six target genes: *wrbA*(Hayashi et al., 2001), which encodes a NADH quinone oxidoreductase; *yehV*
204 (Yokoyama et al., 2000), a transcriptional regulator; *sbcB* (Ohnishi et al., 2002), an exonuclease ;
205 *yecE*, a gene of unknown function; the tRNA gene *argW*(Eppinger et al., 2011a) and Z2577, which
206 encodes an oxidoreductase. Intact reference sequences of these genes were obtained and
207 compared by blastn BLAST(Altschul et al., 1990) against the STEC O157:H7 genome assemblies.
208 Occupied SBI sites were defined as those strains that had disrupted BLAST alignments.

209

210 Clade Typing

211

212 Clade Typing was performed as originally defined by Manning *et al* (2008). The 8 definitive
213 polymorphic positions adopted by Yokoyama *et al* (2012) were used to delineate the strains into the
214 9 clade groupings.

215

216 Locus Specific Polymorphism Assay – LSPA6

217

218 Based on the polymorphic genes defined by Yang *et al* (2004) reference sequences of 6 were
219 extracted from the Sakai reference genome. Sequence alignments were generated using blastn of
220 these sequences against the STEC O157:H7 genome assemblies. The allelic designation '1' was
221 assigned to wild type, '2' assigned to the insertions/deletions defined by Yang *et al* and 'X' to all
222 other polymorphisms.

223

224 *folD-sfmA, Z5935, yhcG, rbsB, rtcB* and *arp-iclR*. Each allele was assigned a number as described
225 previously (Yang et al., 2004). Isolates showing the LSPA6 genotype 111111 were classified as LSPA6
226 lineage I (LSPA6 LI), while those with LSPA6 genotype 211111 were classified as LSPA6 lineage I/II
227 (LSPA6 LI/II). Unique alleles (aberrant amplicon size) were assigned new numbers. All deviations
228 from the genotypes 111111 and 211111 were classified as LSPA6 lineage II (LSPA6 LII).

229

230 Statistical analyses of clinical data amongst clinical cases reported in England

231

The National Enhanced Surveillance System for STEC (NESSS) in England was implemented on 1st January 2009, and has been described in detail elsewhere (*Byrne et al*. 2015, in press). In brief, it collates standardised demographic, clinical and exposure data on all cases of STEC reported in England through collection of a standard enhanced surveillance questionnaire (ESQ). For this study, clinical data on clinical cases for whom strains were sequenced were extracted from NESSS. These data included whether the case reported symptoms of non-bloody diarrhoea; bloody diarrhoea; vomiting; nausea; abdominal pain; fever or whether they were asymptomatic carriers detected through screening high risk contacts of symptomatic cases. Data on whether cases were hospitalised, developed typical HUS or died were also extracted. The age and gender of cases were also extracted. Where clinical symptoms were blank on the ESQ and cases were not recorded as being asymptomatic, these were coded as negative responses. Cases were categorised into children (aged 16 and under) or adults, based on *a priori* knowledge that children are most at risk of both STEC infection and progression to HUS (Byrne et al., 2015). While adults aged over 60 are at increased risk of STEC infection and development of HUS, they were under-represented in these data and were not analysed as a separate group. The outcome of interest was disease severity. Cases were coded as having severe disease if any of the following criteria were reported: Bloody diarrhoea, hospitalisation, HUS or death. Asymptomatic cases and cases with non-bloody diarrhoea were classed as mild.

Genomic variables for analyses included Stx subtype and sublineage. Sublineages were described in respect of Stx subtypes. Cases were described in respect to clinical mild or severe disease and HUS separately) by sublineage. Disease severity was compared amongst gender and age of cases, and sublineage and Fisher's exact tests were used to compare proportions. Logistic Regression analysis was used to investigate phylogenetic groups associated with more severe disease outcomes. Due to the correlation between Stx subtypes and lineage, sublineage was chosen as an explanatory variable for analyses. To assess whether there was a difference in disease severity within sub-lineages they were further subdivided by Stx subtype for analysis. Odds ratios for cases reporting severe disease compared to those reporting mild disease were calculated for each variable. Lineage IIa was chosen as the baseline for lineages as it was found to be the ancestral O157 lineage.

RESULTS

Phylogeny of STEC O157 in the United Kingdom

A maximum likelihood (ML) phylogeny (supplementary figure 1) revealed the population structure of the STEC O157 isolates sequenced in this study. The STEC O157:H7 population has previously been delineated into three lineages, I, I/II and II(Feng et al., 1998; Zhang et al., 2007) and the phylogeny presented here also splits the strains into three groups via deep branches, with reference strains of known lineage(Eppinger et al., 2011b) conforming to the expected pattern.

The ML phylogeny was compared to two other previously used methods to describe the STEC O157 population namely LSPA6 type(Yang et al., 2004) (supplementary figure 1a) and the Manning clade

typing scheme(Manning et al., 2008) (supplementary figure 1b). LSPA6 typing was not congruent with the phylogeny and the lineages defined by LSPA type do not reflect the phylogenetic clustering generated on polymorphisms across the whole genome. By LSPA6 the only strains that type as lineage I (LSPA6 1-1-1-1-1-1) were a clade containing the lineage I strain the assay was designed upon, EDL933. Other strains that cluster within this deep branch (and therefore should be of the same lineage) type as lineage I/II (LSPA6 2-1-1-1-1-1) or had a novel polymorphism. Similarly across the rest of the ML phylogeny the predominant LSPA6 was 2-1-1-1-1-1 or a novel polymorphism. Based on this population, LSPA6 typing did not resolve the lineages correctly and therefore we defined the lineages I, I/II and II based on the deep phylogenetic branches and the placement of reference strains of known lineage.

Supplementary figure 1b shows the phylogeny coloured by clades as described by Manning et al (2008). The clade groupings were broadly congruent with the phylogeny clade 7 (green), clade 8 (purple) and clade 4/5 (cyan) predominated and clade 9 (pink), comprising strains that were β-glucuronidase positive, are an out-group. It was clear however that clade typing does not resolve many phylogenetic splits. In terms of clade typing, lineage II corresponds to clade 7, lineage I/II corresponded to clade 8 and lineage I corresponded to clades 6 through 1 as suggested previously (Eppinger et al., 2011a).

Single linkage clustering based on pairwise genetic distance is an effective method of defining phylogenetic groups as it is inclusive of clonal expansion events. Using a SNP distance threshold of Δ250 we clustered the 1224 strains in this study into 54 groups. 52/54 clusters were distributed within the 3 lineages and there were two outlier clusters, one contained the β-glucuronidase positive strains and another contained 3 isolates associated with travel to Turkey. Supplementary figure 2 shows the number and size of the 52 clusters within the three lineages. Lineage II contained the most diversity with 32 clusters whilst Lineage I and Lineage I/II contained 17 and 3 clusters respectively. All three lineages were associated with uneven sampling of diversity with single high-density clusters comprising 77% of Lineage I isolates, 73% of Lineage I/II isolates and 47% of Lineage II isolates. Isolates contained within the high-density clusters in Lineage I, I/II and II represented the common phage types associated with human infection in the UK: PT21/28, PT2 and PT8 respectively. Isolates in clusters with five or less representatives were more likely to be non-UK strains associated with foreign travel or imported food. Ninety-five isolates were from cattle faecal pats collected as part of a large survey in Scotland(Pearce et al., 2009). These cattle isolates were present in only 8/54 clusters across the three lineages with 84% found in the 3 high-density clusters identified above. This pattern of uneven diversity, coupled with the association of domestic cattle with high-density clones, supports the model of global dispersion and regional expansion of STEC O157:H7.

Recombination

Signals of recombination in the sample population were analysed with BRATNEXTGEN using 270 Δ50 SNP threshold cluster representatives. There were 631,016 recombinant positions found across the 5,498,450 bp alignment and 90% had their origin in the 18 Sakai prophages (SP) or 6 Sakai prophage like elements (SPLE) suggesting that almost all genetic transfer (at least historical) was phage

mediated. The median recombinant size was 575 base pairs whilst the largest was 41212 nucleotides representing an intra-lineage II recombination of SP1. Recombination events were seen at least twice as frequently within lineages (Supplementary table 1) than between lineages with no statistical difference association between the lineage and its likelihood to be a donor or recipient. Within lineage II, the ancestral lineage (see Figure 2) Lineage IIa appeared to be the donor of most recombination events with lineage IIc only receiving foreign DNA. Lineage I had the highest intra-lineage recombination rate, and this that could have contributed to the heterogenous *stx* complement as described in more detail below.

Evolutionary timescale and Stx prophage insertion in STEC O157

A timed phylogeny was constructed using BEAST (Figure 2). The mutation rate of STEC O157:H7 was calculated to be approximately 2.6 mutations/genome/year (95% highest posterior density (HPD) – 2.4 – 2.8) which is in-line with previous estimates for *Escherichia coli*(von Mentzer et al., 2014) and closely related *Shigella* species(Holt et al., 2012). We predict the split of the contemporary β-glucuronidase negative, sorbitol negative clone from the β-glucuronidase positive ancestor to be approximately 400 years ago (95% HPD - 520 years – 301 years). The time to common ancestor of the current circulating diversity (e.g. Lineage I, I/II and II) is approximately 175 years (95% HPD - 198 years – 160 years), significantly more recent than previous estimates of 400 years(Yang et al., 2004) and 2500 years(Leopold et al., 2009). Lineage II is the ancestral lineage which contains at least three sub-lineages that diverged early in the evolutionary process. The most recent common ancestor to Lineage I and Lineage I/II existed approximately 150 years ago (95% HPD - 175 years – 130 years).

The model of Shiga toxin acquisition proposed by Wick and Feng suggested the acquisition of a lambdoid phage containing *stx2* followed by the later acquisition of an *stx1* containing phage (Stx1Φ)(Feng et al., 1998; Wick et al., 2005). The timed phylogeny supported this hypothesis (Figure 2) as the β-glucuronidase positive ancestor and the majority (70%) of stains within lineage IIa and IIb contained only *stx2c*. Sub-lineage Lineage IIc (PT8) (Figure 2) was subsequently lysogenised by an Stx1Φ and had the same disrupted Shiga toxin insertion targets *yehV* and *sbcA* supporting the hypothesis that a truncated prophage was replaced with a Stx1Φ in *yehV*(Shaikh and Tarr, 2003).

The majority of strains in Lineage IIb (PT4/PT1) (Figure 2) carried *stx2c* only but had an occupied *argW* Stx-associated bacteriophage insertion site. There was some further observed heterogeneity in the ancestral lineage IIa with small numbers of dispersed strains containing Stx1Φ, Stx2Φa or being negative for any Shiga toxin alleles as well as having non-stx disrupted stx-associated bacteriophage insertion sites (Supplementary table 2).

The common ancestor of Lineage I/II (Figure 2) was approximately 95 years old marking the divergence of the strain that caused the 2006 Taco Bell outbreak in North America (Sodha et al., 2011) and the PT2 strains associated with the first outbreak of HUS in the United Kingdom in 1983(Taylor et al., 1986). The majority (65%) of strains in lineage I/II were positive for both *stx2c* and

stx2a with occupied SBIs at *yehV* , *sbcA* and *argW*. One sub group of strains belonging to PT2 have subsequently lost Stx2cΦ and had an intact *sbcA* (Supplementary table 3).

Lineage I was by far the most heterogeneous in terms of Stx complement (Supplementary table 4) and arose from a *stx2c*-only ancestor approximately 125 years ago (Figure 2). The majority (87%) of strains in Lineage Ib (PT32) retained the ancestral *stx2c* only genotype of Lineage II and have an additional *yecE* SBI occupied. This lineage had an overrepresentation of strains from Scottish cattle and very few clinical strains. The majority (64%) of strains in Lineage Ia contained Stx2aΦ and Stx1Φ with disrupted *yehV* and *wrbA* including the first fully sequenced STEC O157:H7 genomes (Sakai(Hayashi et al., 2001) and EDL-933(Latif et al., 2014)) and the genome sequence of E. coli O157:H7 strain 2886-75, which was isolated in 1975 making it the oldest STEC O157:H7 strain for which a genome sequence is available (Sanjar et al., 2014). Lineage Ia also contains strains that type as Clade 6 by the Manning scheme and carry the *stx2c* and *stx2a* genes with disrupted *yehV* and *sbcA* which suggests either Stx2aΦ inserted into *yehV* or a novel insertion site.

A final sub-lineage of Lineage I (Lineage Ic) contains 40% of the strains in this study and its common ancestor is approximately 50 years old and has since diverged into 3 clades. These include the ancestral *stx2c* only genotype with occupied *yehV* and *sbcA* SBIs, a *stx2a* only genotype with occupied *yecE*, *yehV* insertion sites and a *stx2a* and *stx2c* genotype with occupied SBIs *yehV*, *sbcA* and *argW*. This final genotype is predominated by phage type 21/28. Within the PT 21/28 clade a sub-clade has subsequently lost the *stx2c* toxin although *yehV*, *sbcA* and *argW* remain occupied.

All 1129 genomes analysed in this study are summarised in terms of Lineage, SNP cluster, SBI, *stx* type, Manning Clade and LSPA-6 type in Supplementary table 5.

Recent Emergence of Predominant UK Lineages

The phage types PT8 and PT21/28 accounted for approximately 60% of clinical isolates identified in the United Kingdom in 2014. Phage typing of STEC O157:H7 in the UK suggests strain replacement has occurred since the beginning of the 21st century with a decline in PT2 corresponding with a rise in PT21/28. PT2 was restricted to lineage I/II whereas PT21/28 was restricted to lineage I indicating strain replacement of one genotype by another distinct genotype, rather than phage type switching within a single genotype.

PT 21/28 typically accounts for >30% of clinical isolates seen in the England, Wales and Scotland each year and is the phage type most commonly associated with outbreaks of HUS(Underwood et al., 2013). As stated above, divergence from the most recent common ancestor occurred 50 years ago subsequently formed into 3 clades; the ancestral PT32 *stx2c* only genotype, a *stx2a* only PT32 genotype associated with travel to Ireland and mainland Europe and finally the PT21/28 clade as a single Δ50 SNP cluster. The PT21/28 clade contained a large number of British cattle (57% of total cattle isolates) and clinical isolates but very few isolates associated with foreign travel (<1%). The

PT21/28 clade arose only 25 years ago and has since undergone a radial expansion resulting in a "comet" like phylogeny (Figure 3.). The PT 21/28 clade itself was flanked by three PT32 *stx2a* and *stx2c* isolates, two from cattle and one clinical isolate from Scotland. It is clear that the direct ancestor of PT21/28 is a PT32 strain.

PT8 was represented as a single Δ250 SNP clonal group (lineage IIc) and its most recent common ancestor can be dated to approximately 50 years ago. Across this clonal group cases were associated with travel to Southern Europe and Northern Africa (22%) suggesting this strain may be endemic in cattle in this region. Within this group there was a recently emerged (30 years to most recent common ancestor) sub-clade where several cases report exposure to domestic cattle, cases report no foreign travel, and there are several strains from UK cattle suggestive of a domestic source of human infection (Figure 4). This again highlights the possibility of imported strains of O157:H7 becoming endemic in local cattle populations.

Disease severity of clinical cases in England by stx subtype and sublineage

A total of 493 strains from clinical cases in England had clinical data available in NESSS. Of those, 311 (63.1%) had experienced bloody diarrhoea, 158 (32.0%) had been hospitalised with their illness and 26 (5.3%) were from cases known to have developed HUS. Thus, two thirds of cases in the dataset were categorised as having severe disease (as defined in methods) however this varied by *stx* subtype and sub-lineage (Table 1). Cases classed as having mild disease accounted for 33.5% of the dataset, and included eighteen asymptomatic cases. Over half (55.4%) of cases in the dataset were female and 55.2% were children aged 16 and under. Severe disease was more frequently reported amongst females (70.3% versus 29.7%, p=0.044) and children (71.9% versus 28.1%, p=0.005).

In univariable analysis, being a child and being female were significantly associated with severe disease (Table 2).All sublineages except Ib and Ic carrying *stx2c*, were significantly associated with more severe disease as compared to sublineage IIa. In the final multivariable model when all variables were controlled for, being a child was a significant predictor of severe disease, but being female was no longer significant. Sub-lineage Ia had the greatest odds of severe disease, with a six-fold increased odds as compared to IIa.

All but one of the HUS cases fell within sub-lineages I-c and I/II (Figure 1) and all were infected with strains carrying *stx2a* either alone or with *stx2c* (Table 2). Lineages Ic and I/II were further divided into strains possessing *stx2a* only and those with *stx2a/2c*. Across all strains, there was no difference in disease severity between cases infected with strains carrying *stx2a* alone or with *2c* (53.5% versus 46.5%, p=0.291). However, in both sublineages Ic and I/II strains carrying *stx2a* only had higher odds of severe disease than those carrying *stx2a/2c* in the final model. While Sub-lineage IIc had increased odds of severe disease, no cases developed HUS. Rather this was due to increased reporting of bloody diarrhoea amongst cases infected with these strains compared to those in other sub-lineages (75.6-% versus 58.6% in other sub-lineages, p=0.005). Most strains (92%) in this sub-lineage carried *stx1a/2c*. Overall, cases infected with strains carrying *stx1a* reported bloody

diarrhoea more frequently than those without (77.5% versus 61.8%, p=0.001) leading to the hypothesis the possession of *stx1a* in strains of sublineage IIc leads to higher rates of bloody diarrhoea.

DISCUSSION

Using phylogenetic analysis of variation at the whole genome level we have been able to reconstruct the phylogenetic history and global diversification of the contemporary STEC O157:H7 clones. The current models of STEC O157:H7 evolution suggest the sero-conversion of an ancestral *stx2 E. coli* O55 to O157. Subsequent loss of the ability to ferment sorbitol and of β-glucuronidase activity gave rise to the common ancestor of the current circulating clone. The evolutionary models of Leopold et al. (2009), Kyle et al.(2012) and Yokoyama et al.(2012) suggest that the β-glucuronidase positive last common ancestor may have given rise to lineage II and lineage I/II in a paraphyletic manner with lineage I/II spawning lineage I (with the acquisition of Stx1 containing lambdoid phage seen in clades 1-3 described by Manning et al. 2008). However, strains had previously been identified that confounded these models and indicated that a more complex explanation was needed (Arthur et al., 2013; Mellor et al., 2013).

In this study we propose a new evolutionary model based on our phylogenetic analysis (Figure 5). In this model we maintain the stepwise series of events from STEC O55 to the β-glucuronidase positive last common ancestor (A5) that evolved into contemporary lineage II. We show at least 3 extant lineages of lineage II including the ancestral branch (IIa) as well as a branch that has acquired Stx1Φ (IIc). A lineage II Stx2cΦ containing strain independently gave rise to Lineage I (approximately 125 years ago) and Lineage I/II (approximately 95 years ago). In lineage I/II a single integration event of a Stx2aΦ into *argW* has been maintained with a sub-group losing Stx2cΦ. Lineage I has a more complex evolutionary history with a Stx2aΦ integrating at least 3 times (once into *wrbA*, once into *argW*, and once into an unknown site), Stx1Φ inserting into lineage Ia strains and at least two loss events of the Stx2cΦ. The model presented here shows StxΦ loss and gain events that have been fixed in the population but we also observe many loss and gain events that appear to be occurring sporadically within each lineage as well as occupation of SBI's with imported DNA that does not encode Stx. This leads to the conclusion that the loss and gain of phage is likely to be highly dynamic but under high selection for retention in the bovine host. Recombination analysis highlighted the phage regions to be hotspots of DNA exchange, with remarkably little activity outside these regions.

In this analysis we predict the split from the β-glucuronidase positive last common ancestor (A5) to have occurred approximately 400 years ago with the common ancestor of the current diversity appearing 175 years ago. At this point there was an expansion event with the major lineages formed within 30 or so years. This early diversification of STEC O157:H7 fits with the extant diversity of STEC O157:H7 being globally distributed. Although a large degree of diversity of STEC O157:H7 is seen in the UK, the distribution of this diversity is uneven. We show that several pockets of diversity are seen at much higher frequency then others and that the same pockets of diversity are more

frequently observed in both human clinical cases and in the local cattle population. This fits with model of historical dissemination of diversity and then regional expansion in native cattle with occasional sampling of the wider diversity through imported foodstuff and foreign travel.

Although we have shown the contemporary clone existed over 100 years earlier, STEC O157:H7 only became a recognised pathogen in the 1980's(Riley et al., 1983) after causing outbreaks of severe illness. Whilst STEC O157:H7 causes gastroenteritis in most infections a significant minority develop more severe symptoms including HUS. Whilst progression to HUS no doubt has many host predictors, a clear association with the presence of *stx2a* subtype has been shown(Persson et al., 2007). In our study we show that the acquisition of the *stx2a* subtype occurred relatively recently compared to the other stx subtypes and is likely to explain the recent emergence of the STEC O157:H7 serotype as a clinically significant pathogen. We also show that *stx2a* is likely to have been acquired by STEC O157:H7 on multiple occasions highlighting the potential for new, highly virulent clones to emerge. Finally it appears that once *stx2a* is integrated in a population it tends to be maintained, often at the expense of *stx2c*. Recent research has indicated that the Stx2aΦ is associated not only with more severe human disease but also with higher excretion levels in cattle(Matthews et al., 2013).

Using clinical outcome data on a cohort of nearly 500 STEC O157:H7 cases we are able to assess the risk of serve disease of each of the extant lineages and sub-lineages. The presence of *stx2a* is a pre-requisite for the development of HUS with 100% of HUS cases infected with a strain harbouring this toxin sub-type. Multivariable regression analysis with the ancestral IIa clone as the baseline shows IIc has a nearly 4-fold increase in risk of severe disease accounted by in increase in incidence of bloody diarrhea. This PT8 clone has acquired a Stx1Φ carrying the same Stx as found in *Shigella dysenteriae* serotype 1. All sub-lineages of lineage I and I/II that contain *stx2a* have an increased risk of severe disease with the additional presence of *stx2c* appearing to have a protective effect. This presumably reflects regulatory interactions between the prophages. These analyses show the clear importance of determining the Stx complement of an STEC O157 strain when predicting the likely risk of severe disease and therefore case management.

This study shows that recent strain replacement has occurred in Great Britain shaping the diversity of STEC O157:H7 observed today. Within lineage II, an importation of a PT8 strain probably from the Mediterranean cattle population of Southern Europe and Northern Africa occurred within the last 30 years. Similarly within the last 25 years the emergence and rapid expansion of PT 21/28 in lineage I in Great Britain led to this highly virulent subtype being found ubiquitously in domestic cattle. These recent strain replacement events provide insight into the dynamics of STEC O157:H7 transmission on a national and international scale and suggest that while the overall diversity of this pathogen is globally distributed, regionally endemic strains can be transmitted and eventually become the dominant strain in the local cattle population. Whilst the imported strain may play a role in out-competing domestic strains, agricultural practices such as culling and restocking of animals, as seen during the foot and mouth disease and Bovine Spongiform Encephalitis (BSE) epidemics may act as drivers facilitating more rapid strain replacement (Carrique-Mas et al., 2008).

From the current study it appears the relatively high incidence of STEC O157 human infections in the UK results from the emergence and expansion of a Lineage I PT21/28 clade in the last 25 years, producing strains containing both Stx2a and Stx2c prophages that are capable of higher excretion levels from cattle (super-shedding) and can cause severe disease in humans. Therefore, screening and intervention strategies should be targeting these strain clusters that are the most significant threat to human health. Further work is needed to understand the diversity of host phages that carry Stx and the reasons behind the proliferation of this cluster. While Stx is essential for the severe pathology associated with human STEC disease, the role of the different toxins in governing supershedding is unknown. Moreover, it is evident that other genes on Stx-encoding prophages regulate the expression of bacterial colonisation factors and this will also impact on the success of the cluster(Xu et al., 2012).

ACKNOWLEDGEMENTS

This work was funded by the National Institute for Health Research scientific research development fund (108601). Food Standards Agency programme FS101055 and a BBSRC Institute Strategic Programme to the Roslin Institute.

ABBREVIATIONS

REFERENCES

Abu-Ali, G.S., Ouellette, L.M., Henderson, S.T., Lacher, D.W., Riordan, J.T., Whittam, T.S., Manning, S.D., 2010. Increased Adherence and Expression of Virulence Genes in a Lineage of Escherichia coli O157:H7 Commonly Associated with Human Infections. PLoS ONE 5, e10167. doi:10.1371/journal.pone.0010167

Ahmed, R., Bopp, C., Borczyk, A., Kasatiya, S., 1987. Phage-typing scheme for Escherichia coli O157:H7. J. Infect. Dis. 155, 806–809.

Akashi, S., Joh, K., Tsuji, A., Ito, H., Hoshi, H., Hayakawa, T., Ihara, J., Abe, T., Hatori, M., Mori, T., 1994. A severe outbreak of haemorrhagic colitis and haemolytic uraemic syndrome associated with Escherichia coli O157:H7 in Japan. Eur. J. Pediatr. 153, 650–655.

563 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool.
564 J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

565 Arthur, T.M., Ahmed, R., Chase-Topping, M., Kalchayanand, N., Schmidt, J.W., Bono, J.L., 2013.
566 Characterization of Escherichia coli O157:H7 Strains Isolated from Supershedding Cattle. Appl.
567 Environ. Microbiol. 79, 4294–4303. doi:10.1128/AEM.00846-13

568 Ashton, P.M., Perry, N., Ellis, R., Petrovska, L., Wain, J., Grant, K.A., Jenkins, C., Dallman, T.J., 2015.
569 Insight into Shiga toxin genes encoded by *Escherichia coli* O157 from whole genome sequencing.
570 PeerJ 3, e739. doi:10.7717/peerj.739

571 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko,
572 S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A.,
573 Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell
574 sequencing. J. Comput. Biol. J. Comput. Mol. Cell Biol. 19, 455–477. doi:10.1089/cmb.2012.0021

575 Besser, T.E., Shaikh, N., Holt, N.J., Tarr, P.I., Konkel, M.E., Malik-Kale, P., Walsh, C.W., Whittam, T.S.,
576 Bono, J.L., 2007. Greater Diversity of Shiga Toxin-Encoding Bacteriophage Insertion Sites among
577 Escherichia coli O157:H7 Isolates from Cattle than in Those from Humans. Appl. Environ. Microbiol.
578 73, 671–679. doi:10.1128/AEM.01035-06

579 Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence
580 data. Bioinforma. Oxf. Engl. 30, 2114–2120. doi:10.1093/bioinformatics/btu170

581 Bono, J.L., Keen, J.E., Clawson, M.L., Durso, L.M., Heaton, M.P., Laegreid, W.W., 2007. Association of
582 Escherichia coli O157:H7 tir polymorphisms with human infection. BMC Infect. Dis. 7, 98.
583 doi:10.1186/1471-2334-7-98

584 Byrne, L., Jenkins, C., Launders, N., Elson, R., Adak, G.K., 2015. The epidemiology, microbiology and
585 clinical impact of Shiga toxin-producing Escherichia coli in England, 2009-2012. Epidemiol. Infect. 1–
586 13. doi:10.1017/S0950268815000746

587 Carrique-Mas, J.J., Medley, G.F., Green, L.E., 2008. Risks for bovine tuberculosis in British cattle
588 farms restocked after the foot and mouth disease epidemic of 2001. Prev. Vet. Med. 84, 85–93.
589 doi:10.1016/j.prevetmed.2007.11.001

590 Centers for Disease Control and Prevention (CDC), 2006. Ongoing multistate outbreak of Escherichia
591 coli serotype O157:H7 infections associated with consumption of fresh spinach--United States,
592 September 2006. MMWR Morb. Mortal. Wkly. Rep. 55, 1045–1046.

593 Chase-Topping, M., Gally, D., Low, C., Matthews, L., Woolhouse, M., 2008. Super-shedding and the
594 link between human infection and livestock carriage of Escherichia coli O157. Nat. Rev. Microbiol. 6,
595 904–912. doi:10.1038/nrmicro2029

596 Dowd, S.E., Williams, J.B., 2008. Comparison of Shiga-like toxin II expression between two genetically
597 diverse lineages of Escherichia coli O157:H7. J. Food Prot. 71, 1673–1678.

598 Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and
599 the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973. doi:10.1093/molbev/mss075

600 Eppinger, M., Mammel, M.K., Leclerc, J.E., Ravel, J., Cebula, T.A., 2011a. Genomic anatomy of
601 Escherichia coli O157:H7 outbreaks. Proc. Natl. Acad. Sci. 108, 20142–20147.
602 doi:10.1073/pnas.1107176108

603 Eppinger, M., Mammel, M.K., LeClerc, J.E., Ravel, J., Cebula, T.A., 2011b. Genome Signatures of
604 Escherichia coli O157:H7 Isolates from the Bovine Host Reservoir. Appl. Environ. Microbiol. 77,
605 2916–2925. doi:10.1128/AEM.02554-10

606 Feng, P., Lampel, K.A., Karch, H., Whittam, T.S., 1998. Genotypic and Phenotypic Changes in the
607 Emergence of Escherichia coli O157:H7. J. Infect. Dis. 177, 1750–1753. doi:10.1086/517438

608 Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.-G., Ohtsubo, E.,
609 Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C.,
610 Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., Shinagawa, H., 2001. Complete
611 Genome Sequence of Enterohemorrhagic Eschelichia coli O157:H7 and Genomic Comparison with a
612 Laboratory Strain K-12. DNA Res. 8, 11–22. doi:10.1093/dnares/8.1.11

613 Holt, K.E., Baker, S., Weill, F.-X., Holmes, E.C., Kitchen, A., Yu, J., Sangal, V., Brown, D.J., Coia, J.E.,
614 Kim, D.W., Choi, S.Y., Kim, S.H., da Silveira, W.D., Pickard, D.J., Farrar, J.J., Parkhill, J., Dougan, G.,
615 Thomson, N.R., 2012. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent
616 global dissemination from Europe. Nat. Genet. 44, 1056–1059. doi:10.1038/ng.2369

617 Ihekweazu, C., Carroll, K., Adak, B., Smith, G., Pritchard, G.C., Gillespie, I.A., Verlander, N.Q., Harvey-
618 Vince, L., Reacher, M., Edeghere, O., Sultan, B., Cooper, R., Morgan, G., Kinross, P.T.N., Boxall, N.S.,
619 Iversen, A., Bickler, G., 2012. Large outbreak of verocytotoxin-producing Escherichia coli O157
620 infection in visitors to a petting farm in South East England, 2009. Epidemiol. Infect. 140, 1400–1413.
621 doi:10.1017/S0950268811002111

622 Khakhria, R., Duck, D., Lior, H., 1990. Extended phage-typing scheme for Escherichia coli O157:H7.
623 Epidemiol. Infect. 105, 511–520.

624 Kim, J., Nietfeldt, J., Ju, J., Wise, J., Fegan, N., Desmarchelier, P., Benson, A.K., 2001. Ancestral
625 Divergence, Genome Diversification, and Phylogeographic Variation in Subpopulations of Sorbitol-
626 Negative, ?-Glucuronidase-Negative Enterohemorrhagic Escherichia coli O157. J. Bacteriol. 183,
627 6885–6897. doi:10.1128/JB.183.23.6885-6897.2001

628 Kyle, J.L., Cummings, C.A., Parker, C.T., Quiñones, B., Vatta, P., Newton, E., Huynh, S., Swimley, M.,
629 Degoricija, L., Barker, M., Fontanoz, S., Nguyen, K., Patel, R., Fang, R., Tebbs, R., Petrauskene, O.,
630 Furtado, M., Mandrell, R.E., 2012. Escherichia coli Serotype O55:H7 Diversity Supports Parallel
631 Acquisition of Bacteriophage at Shiga Toxin Phage Insertion Sites during Evolution of the O157:H7
632 Lineage. J. Bacteriol. 194, 1885–1896. doi:10.1128/JB.00120-12

633 Lai, Y., Rosenshine, I., Leong, J.M., Frankel, G., 2013. Intimate host attachment: enteropathogenic
634 and enterohaemorrhagic Escherichia coli. Cell. Microbiol. 15, 1796–1808. doi:10.1111/cmi.12179

635 Latif, H., Li, H.J., Charusanti, P., Palsson, B.Ø., Aziz, R.K., 2014. A Gapless, Unambiguous Genome
636 Sequence of the Enterohemorrhagic Escherichia coli O157:H7 Strain EDL933. Genome Announc. 2.
637 doi:10.1128/genomeA.00821-14

638 Lee, K., French, N.P., Jones, G., Hara-Kudo, Y., Iyoda, S., Kobayashi, H., Sugita-Konishi, Y., Tsubone,
639 H., Kumagai, S., 2012. Variation in Stress Resistance Patterns among stx Genotypes and Genetic
640 Lineages of Shiga Toxin-Producing Escherichia coli O157. Appl. Environ. Microbiol. 78, 3361–3368.
641 doi:10.1128/AEM.06646-11

642 Leopold, S.R., Magrini, V., Holt, N.J., Shaikh, N., Mardis, E.R., Cagno, J., Ogura, Y., Iguchi, A., Hayashi,
643 T., Mellmann, A., Karch, H., Besser, T.E., Sawyer, S.A., Whittam, T.S., Tarr, P.I., 2009. A precise
644 reconstruction of the emergence and constrained radiations of Escherichia coli O157 portrayed by

backbone concatenomic analysis. Proc. Natl. Acad. Sci. 106, 8713–8718. doi:10.1073/pnas.0812949106

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 26, 589–595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Manning, S.D., Motiwala, A.S., Springman, A.C., Qi, W., Lacher, D.W., Ouellette, L.M., Mladonicky, J.M., Somsel, P., Rudrik, J.T., Dietrich, S.E., Zhang, W., Swaminathan, B., Alland, D., Whittam, T.S., 2008. Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. Proc. Natl. Acad. Sci. 105, 4868–4873. doi:10.1073/pnas.0710834105

Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D., Corander, J., 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 40, e6. doi:10.1093/nar/gkr928

Matthews, L., Reeve, R., Gally, D.L., Low, J.C., Woolhouse, M.E.J., McAteer, S.P., Locking, M.E., Chase-Topping, M.E., Haydon, D.T., Allison, L.J., Hanson, M.F., Gunn, G.J., Reid, S.W.J., 2013. Predicting the public health benefit of vaccinating cattle against Escherichia coli O157. Proc. Natl. Acad. Sci. U. S. A. 110, 16265–16270. doi:10.1073/pnas.1304978110

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. doi:10.1101/gr.107524.110

Mellor, G.E., Besser, T.E., Davis, M.A., Beavis, B., Jung, W., Smith, H.V., Jennison, A.V., Doyle, C.J., Chandry, P.S., Gobius, K.S., Fegan, N., 2013. Multilocus Genotype Analysis of Escherichia coli O157 Isolates from Australia and the United States Provides Evidence of Geographic Divergence. Appl. Environ. Microbiol. 79, 5050–5058. doi:10.1128/AEM.01525-13

Ohnishi, M., Terajima, J., Kurokawa, K., Nakayama, K., Murata, T., Tamura, K., Ogura, Y., Watanabe, H., Hayashi, T., 2002. Genomic diversity of enterohemorrhagic Escherichia coli O157 revealed by whole genome PCR scanning. Proc. Natl. Acad. Sci. 99, 17043–17048. doi:10.1073/pnas.262441699

Pearce, M.C., Chase-Topping, M.E., McKendrick, I.J., Mellor, D.J., Locking, M.E., Allison, L., Ternent, H.E., Matthews, L., Knight, H.I., Smith, A.W., Synge, B.A., Reilly, W., Low, J.C., Reid, S.W.J., Gunn, G.J., Woolhouse, M.E.J., 2009. Temporal and spatial patterns of bovine Escherichia coli O157 prevalence and comparison of temporal changes in the patterns of phage types associated with bovine shedding and human E. coli O157 cases in Scotland between 1998-2000 and 2002-2004. BMC Microbiol. 9, 276. doi:10.1186/1471-2180-9-276

Persson, S., Olsen, K.E.P., Ethelberg, S., Scheutz, F., 2007. Subtyping method for Escherichia coli shiga toxin (verocytotoxin) 2 variants and correlations to clinical manifestations. J. Clin. Microbiol. 45, 2020–2024. doi:10.1128/JCM.02591-06

Riley, L.W., Remis, R.S., Helgerson, S.D., McGee, H.B., Wells, J.G., Davis, B.R., Hebert, R.J., Olcott, E.S., Johnson, L.M., Hargrett, N.T., Blake, P.A., Cohen, M.L., 1983. Hemorrhagic colitis associated with a rare Escherichia coli serotype. N. Engl. J. Med. 308, 681–685. doi:10.1056/NEJM198303243081203

686 Riordan, J.T., Viswanath, S.B., Manning, S.D., Whittam, T.S., 2008. Genetic Differentiation of
687 Escherichia coli O157:H7 Clades Associated with Human Disease by Real-Time PCR. J. Clin. Microbiol.
688 46, 2070–2073. doi:10.1128/JCM.00203-08

689 Sanjar, F., Hazen, T.H., Shah, S.M., Koenig, S.S.K., Agrawal, S., Daugherty, S., Sadzewicz, L., Tallon, L.J.,
690 Mammel, M.K., Feng, P., Soderlund, R., Tarr, P.I., DebRoy, C., Dudley, E.G., Cebula, T.A., Ravel, J.,
691 Fraser, C.M., Rasko, D.A., Eppinger, M., 2014. Genome Sequence of Escherichia coli O157:H7 Strain
692 2886-75, Associated with the First Reported Case of Human Infection in the United States. Genome
693 Announc. 2, e01120–13. doi:10.1128/genomeA.01120-13

694 Schmidt, H., Karch, H., Beutin, L., 1994. The large-sized plasmids of enterohemorrhagic Escherichia
695 coli O157 strains encode hemolysins which are presumably members of the E. coli alpha-hemolysin
696 family. FEMS Microbiol. Lett. 117, 189–196.

697 Scotland, S.M., Smith, H.R., Rowe, B., 1985. Two distinct toxins active on Vero cells from Escherichia
698 coli O157. Lancet 2, 885–886.

699 Shaikh, N., Tarr, P.I., 2003. Escherichia coli O157:H7 Shiga Toxin-Encoding Bacteriophages:
700 Integrations, Excisions, Truncations, and Evolutionary Implications. J. Bacteriol. 185, 3596–3605.
701 doi:10.1128/JB.185.12.3596-3605.2003

702 Sodha, S.V., Lynch, M., Wannemuehler, K., Leeper, M., Malavet, M., Schaffzin, J., Chen, T., Langer, A.,
703 Glenshaw, M., Hoefer, D., Dumas, N., Lind, L., Iwamoto, M., Ayers, T., Nguyen, T., Biggerstaff, M.,
704 Olson, C., Sheth, A., Braden, C., 2011. Multistate outbreak of Escherichia coli O157:H7 infections
705 associated with a national fast-food chain, 2006: a study incorporating epidemiological and food
706 source traceback results. Epidemiol. Infect. 139, 309–316. doi:10.1017/S0950268810000920

707 Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
708 phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

709 Taylor, C.M., White, R.H., Winterborn, M.H., Rowe, B., 1986. Haemolytic-uraemic syndrome: clinical
710 experience of an outbreak in the West Midlands. Br. Med. J. Clin. Res. Ed 292, 1513–1516.

711 Tobe, T., Beatson, S.A., Taniguchi, H., Abe, H., Bailey, C.M., Fivian, A., Younis, R., Matthews, S.,
712 Marches, O., Frankel, G., Hayashi, T., Pallen, M.J., 2006. An extensive repertoire of type III secretion
713 effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. Proc. Natl.
714 Acad. Sci. U. S. A. 103, 14941–14946. doi:10.1073/pnas.0604891103

715 Underwood, A.P., Dallman, T., Thomson, N.R., Williams, M., Harker, K., Perry, N., Adak, B., Willshaw,
716 G., Cheasty, T., Green, J., Dougan, G., Parkhill, J., Wain, J., 2013. Public Health Value of Next-
717 Generation DNA Sequencing of Enterohemorrhagic Escherichia coli Isolates from an Outbreak. J.
718 Clin. Microbiol. 51, 232–237. doi:10.1128/JCM.01696-12

719 Von Mentzer, A., Connor, T.R., Wieler, L.H., Semmler, T., Iguchi, A., Thomson, N.R., Rasko, D.A.,
720 Joffre, E., Corander, J., Pickard, D., Wiklund, G., Svennerholm, A.-M., Sjöling, A., Dougan, G., 2014.
721 Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution.
722 Nat. Genet. 46, 1321–1326. doi:10.1038/ng.3145

723 Whittam, T.S., Wachsmuth, I.K., Wilson, R.A., 1988. Genetic evidence of clonal descent of Escherichia
724 coli O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome. J. Infect. Dis. 157,
725 1124–1133.

Wick, L.M., Qi, W., Lacher, D.W., Whittam, T.S., 2005. Evolution of Genomic Content in the Stepwise Emergence of Escherichia coli O157:H7. J. Bacteriol. 187, 1783–1791. doi:10.1128/JB.187.5.1783-1791.2005

Xu, X., McAteer, S.P., Tree, J.J., Shaw, D.J., Wolfson, E.B.K., Beatson, S.A., Roe, A.J., Allison, L.J., Chase-Topping, M.E., Mahajan, A., Tozzoli, R., Woolhouse, M.E.J., Morabito, S., Gally, D.L., 2012. Lysogeny with Shiga toxin 2-encoding bacteriophages represses type III secretion in enterohemorrhagic Escherichia coli. PLoS Pathog. 8, e1002672. doi:10.1371/journal.ppat.1002672

Yang, Z., Kovar, J., Kim, J., Nietfeldt, J., Smith, D.R., Moxley, R.A., Olson, M.E., Fey, P.D., Benson, A.K., 2004. Identification of Common Subpopulations of Non-Sorbitol-Fermenting, β-Glucuronidase-Negative Escherichia coli O157:H7 from Bovine Production Environments and Human Clinical Samples. Appl. Environ. Microbiol. 70, 6846–6854. doi:10.1128/AEM.70.11.6846-6854.2004

Yokoyama, E., Hirai, S., Hashimoto, R., Uchimura, M., 2012. Clade analysis of enterohemorrhagic Escherichia coli serotype O157:H7/H- strains and hierarchy of their phylogenetic relationships. Infect. Genet. Evol. 12, 1724–1728. doi:10.1016/j.meegid.2012.07.003

Yokoyama, K., Makino, K., Kubota, Y., Watanabe, M., Kimura, S., Yutsudo, C.H., Kurokawa, K., Ishii, K., Hattori, M., Tatsuno, I., Abe, H., Yoh, M., Iida, T., Ohnishi, M., Hayashi, T., Yasunaga, T., Honda, T., Sasakawa, C., Shinagawa, H., 2000. Complete nucleotide sequence of the prophage VT1-Sakai carrying the Shiga toxin 1 genes of the enterohemorrhagic Escherichia coli O157:H7 strain derived from the Sakai outbreak. Gene 258, 127–139.

Zhang, Y., Laing, C., Steele, M., Ziebell, K., Johnson, R., Benson, A.K., Taboada, E., Gannon, V.P., 2007. Genome evolution in major Escherichia coli O157:H7 lineages. BMC Genomics 8, 121. doi:10.1186/1471-2164-8-121

DATA BIBLIOGRAPHY

1. Dallman, T. J., Ashton, P. A., Jenkins, C., Grant K. NCBI Short Read Archive. PRJNA248042 (2015).

2. Dallman, T. J. GitHub https://github.com/timdallman/phylogenomics_stec (2015).

FIGURES AND TABLES

Figure 1. Proportion of cases of the predominant phage types in England & Wales and Scotland over the last 20 years.

Figure 2. Maximum clade credibility tree of 530 Δ25 SNP representatives. The tree is highlighted by lineage and the loss and gain of StxΦ with the associated Stx-associated bacteriophage insertion (SBI) in brackets. The GUD+ lineage represents the strains that retained the ability to express β-glucuronidase.

Figure 3. Left - maximum likelihood phylogeny of 400 lineage I Δ5 SNP representatives with lineage Ic highlighted in grey. Right – maximum likelihood phylogeny of lineage Ic showing the radial expansion of PT21/28 from the PT32 ancestor with isolates annotated by cattle or clinical origin.

Figure 4. Left - maximum likelihood phylogeny of 241 lineage II Δ5 SNP representatives with lineage IIc (PT8) highlighted in grey. Right – maximum likelihood phylogeny of lineage IIc showing the distribution of Mediterranean travel associated cases and UK cattle cases.

Figure 5. STEC O157:H7 evolutionary model based on a timed phylogeny of over 1000 genomes showing the key evolutionary splits and the associated gain and loss of stx containing prophage. GUD+ represents strains that have the ability to express β-glucuronidase, sor+ represents strains that have the ability to ferment sorbital.

Sublineage	Mild		Severe[1]		Totals		%HUS[2]	
	n	%	n	%	n	%	n	%
II a	42	56.8	32	43.2	74	100	1	1.4%
II b	18	81.8	4	18.2	22	100	0	0.0%
II c	31	23.7	100	76.3	131	100	1	0.8%
I a	3	17.7	14	82.3	17	100	0	0.0%
I b	7	77.8	2	22.2	9	100	0	0.0%
Ic (stx2a)	9	20.9	34	79.1	43	100	8	18.6%
Ic (stx 2a/2c)	35	30.2	81	69.8	116	100	10	8.6%
Ic (stx2c)	1	25	3	75.0	4	100	0	0.0%
I/II (stx2a)	7	18.4	31	81.6	38	100	2	5.3%
I/II (stx2a/2c)	12	30.8	27	69.2	39	100	4	10.3%
All strains	165	33.5	328	66.5	493	100	26	5.3%

Table 1:

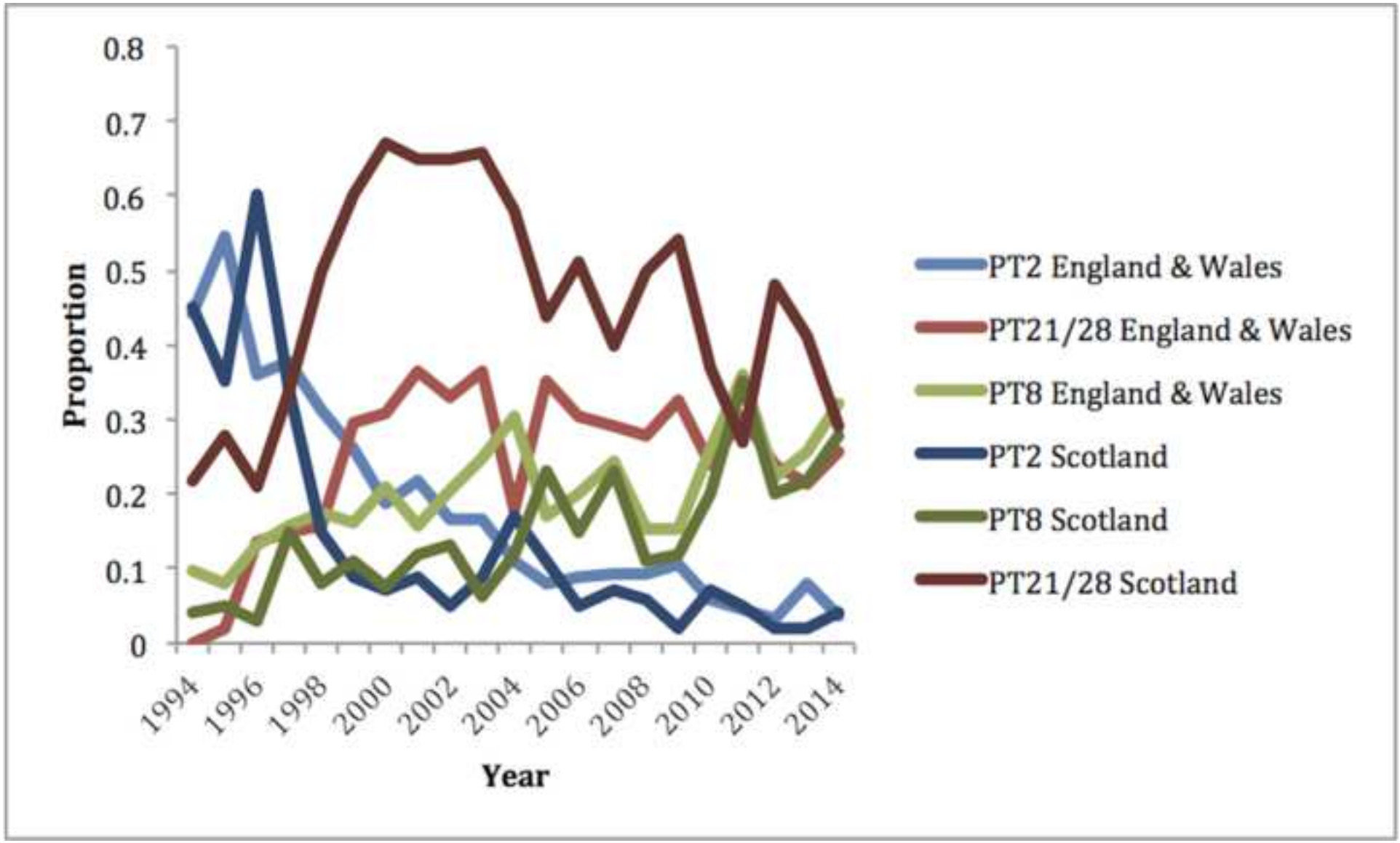
Sub-lineage and *stx* subtype of whole genome sequenced strains isolated from clinical cases of STEC O157 in England. [1]Includes cases with bloody diarrhoea or cases who were hospitalised. [2]The lineage IIa strain isolated from a patient with HUS possessed *stx2a/2c*; The lineage IIc strain possessed *stx1a/2a/2c*.

Univariable					
Variable	**Category**	**Odds Ratio**	**P-value**	**Lower 95% CI**	**Upper 95% CI**
Age	*Child*	*1.73*	*0.005*	*1.18*	*2.51*
	Adult	*Baseline*			
Sex	*Female*	*1.49*	*0.037*	*1.02*	*2.17*
	Male	*Baseline*			
Sub lineage	*II a*	*Baseline*			
	II b	*0.29*	*0.040*	*0.09*	*0.95*
	II c	*4.23*	*0.000*	*2.30*	*7.80*
	I a	*6.12*	*0.008*	*1.62*	*23.14*
	I b	*0.37*	*0.240*	*0.07*	*1.93*
	Ic (stx2a)	*4.96*	*<0.001*	*2.08*	*11.80*
	Ic (stx2a/2c)	*2.92*	*0.001*	*1.59*	*5.34*
	Ic (stx2c)	*3.94*	*0.245*	*0.39*	*39.65*
	I/II (stx2a)	*5.81*	*<0.001*	*2.27*	*14.88*
	I/II (stx2a/2c)	*2.95*	*0.010*	*1.30*	*6.71*
Multivariable Analysis					
Variable	**Category**	**Odds Ratio**	**P-value**	**Lower 95% CI**	**Upper 95% CI**
Age	*Child*	*1.56*	*0.042*	*1.01*	*2.39*

	Adult	*Baseline*			
Sex	*Female*	*1.15*	*0.489*	*0.76*	*1.75*
	Male	*Baseline*			
Sub lineage	*II a*	*Baseline*			
	II b	*0.29*	*0.040*	*0.09*	*0.95*
	II c	*3.65*	*<0.001*	*1.95*	*6.83*
	I a	*6.09*	*0.008*	*1.60*	*23.20*
	I b	*0.35*	*0.209*	*0.67*	*1.81*
	Ic (stx2a)	*5.05*	*<0.001*	*2.11*	*12.10*
	Ic (stx2a/2c)	*3.06*	*<0.001*	*1.66*	*5.67*
	Ic (stx2c)	*3.48*	*0.293*	*0.34*	*35.62*
	I/II (stx2a)	*4.89*	*0.001*	*1.88*	*12.73*
	I/II stx(stx2a/2c)	*2.87*	*0.012*	*1.26*	*6.58*

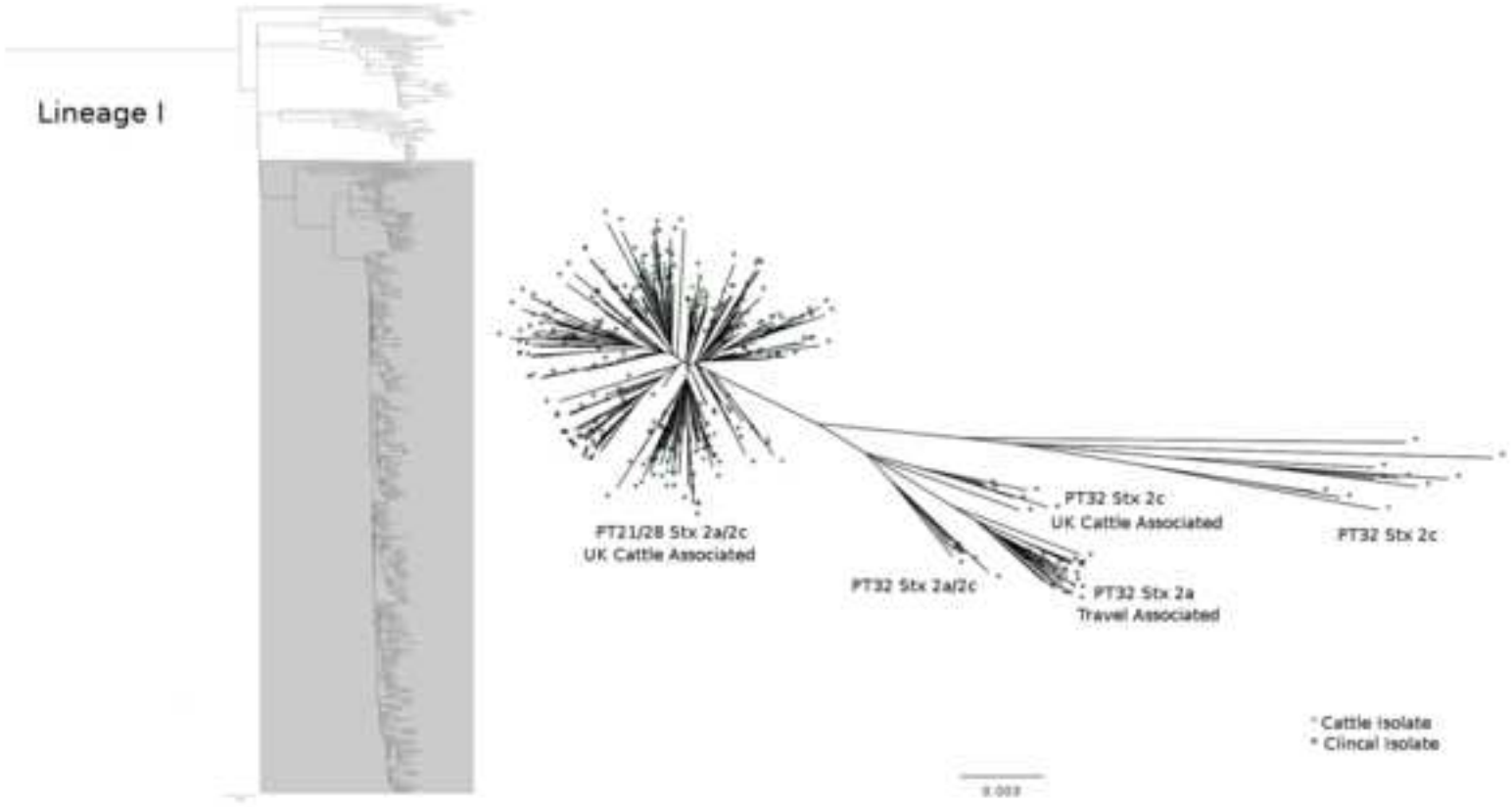
Table 2:

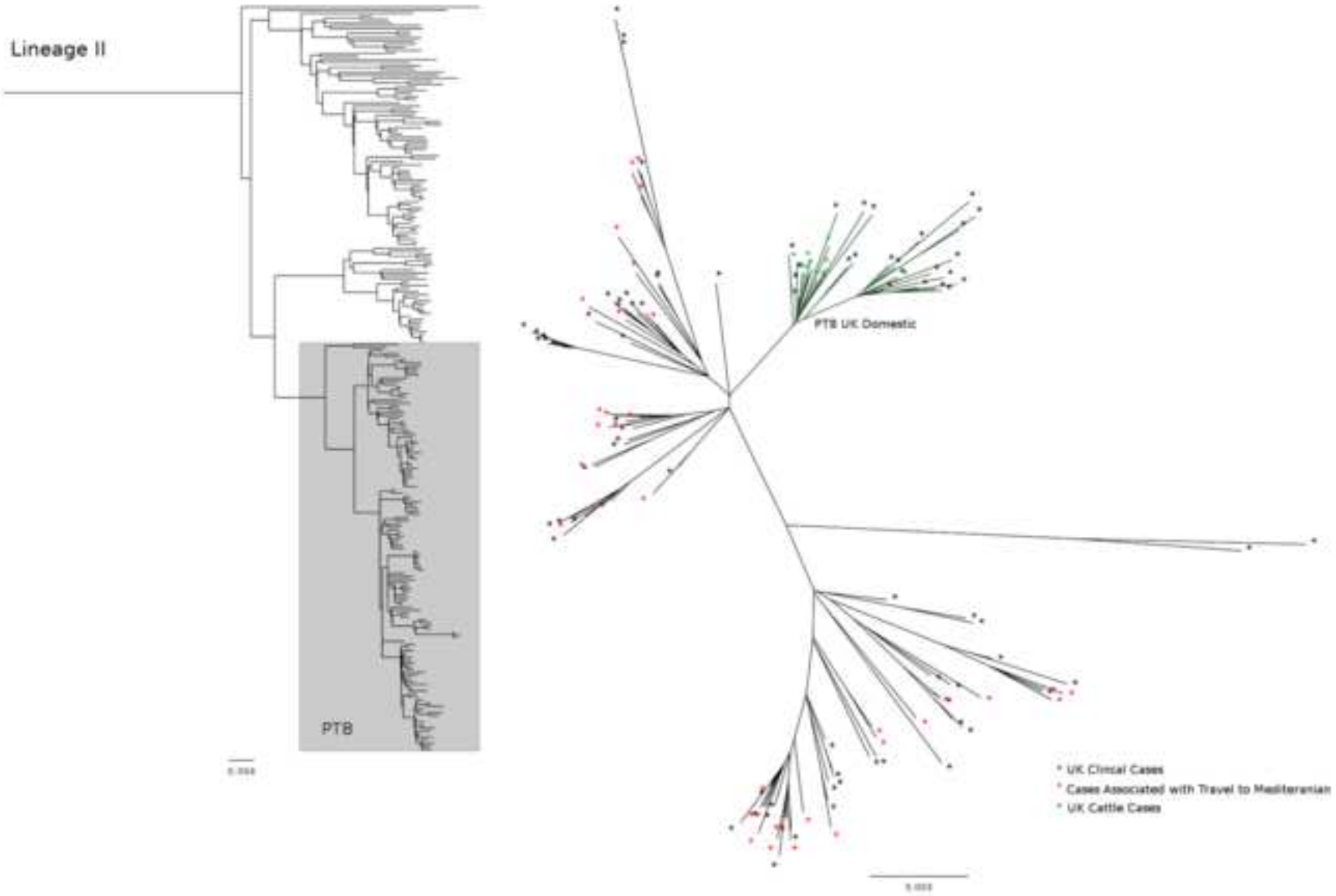
Disease severity amongst clinical cases of STEC O157 in England where strains had been whole genome sequenced by age, gender and sublineage.



Click here to download Figure: Figure2.tiff

Click here to download Figure: Figure3.tiff





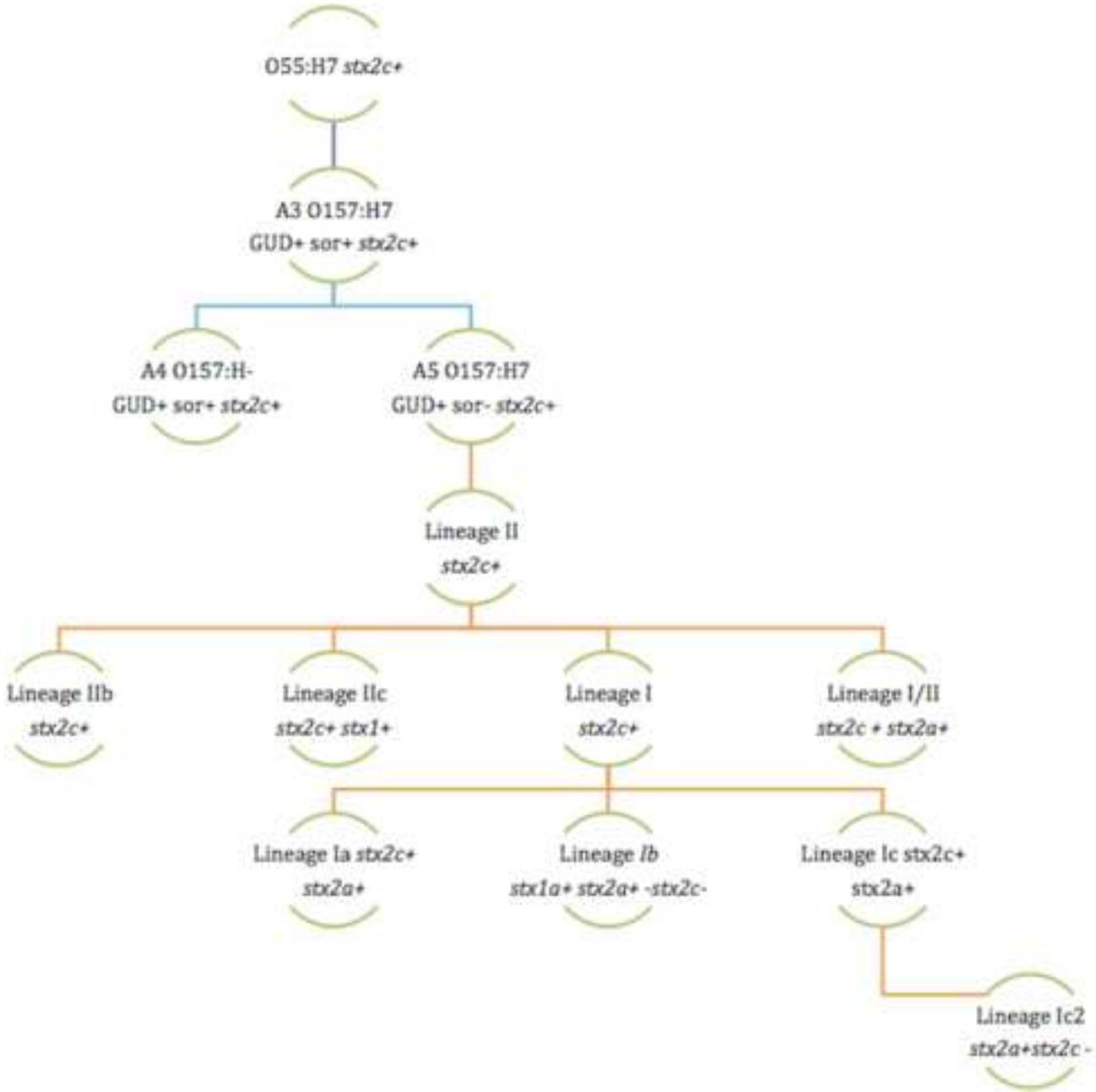
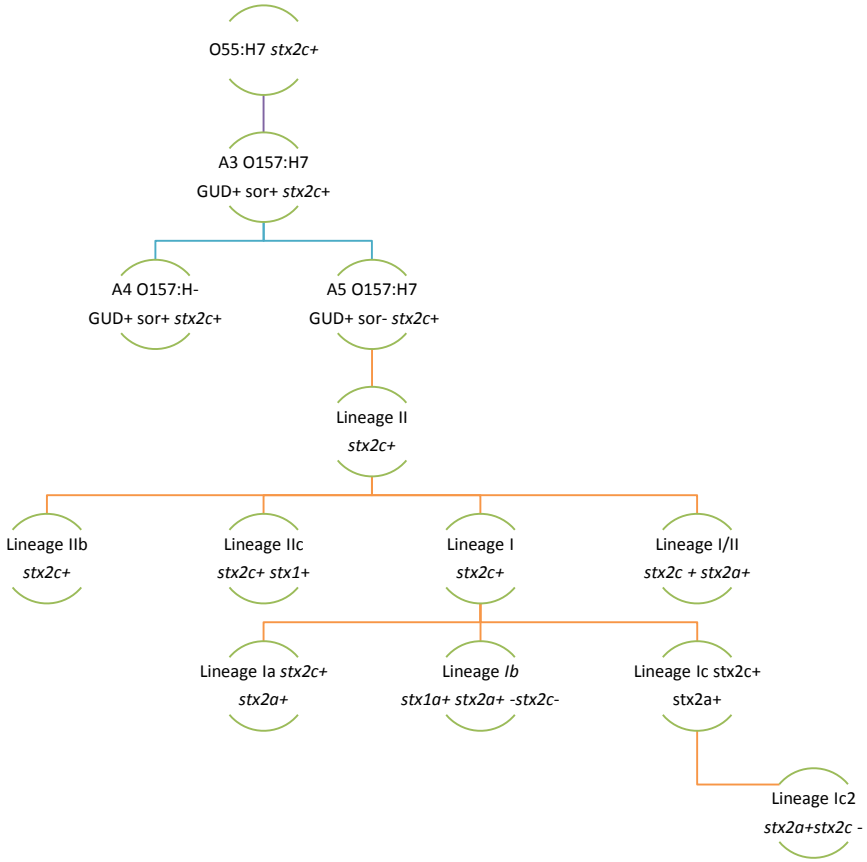


Figure 5 as Powerpoint

Click here to download Figure: Figure5.pptx

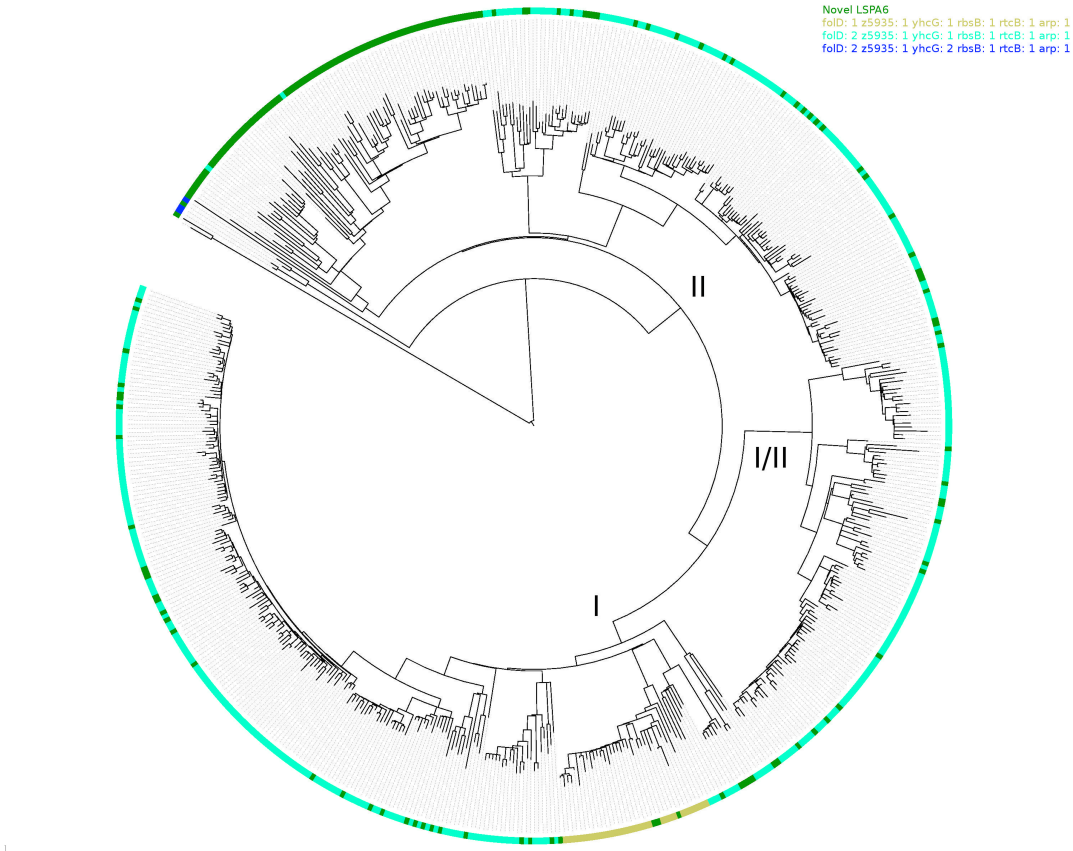


Phylogenomics of Shiga Toxin producing *Escherichia coli* O157:H7: assessing the risk of severe human disease in light of recent strain replacement in the cattle population in the United Kingdom

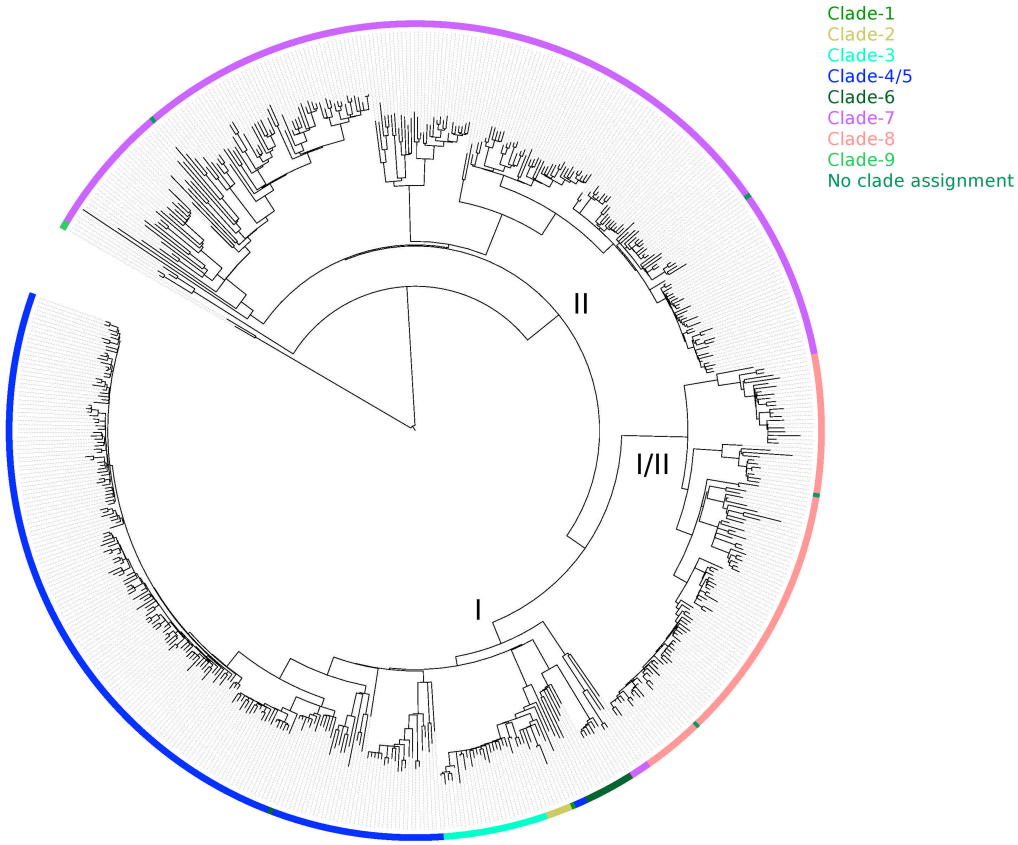
Timothy J. Dallman[1*], Philip M. Ashton[1], Lisa Byrne[1], Neil T. Perry[1], Liljana Petrovska[3], Richard Ellis[3], Lesley Allison[5], Mary Hanson[5], Anne Holmes[5], George J. Gunn[7], Margo E. Chase-Topping[6], Mark E. J. Woolhouse[6], Kathie A. Grant[1], David L. Gally[4], John Wain[2*], Claire Jenkins[1].

[1]Public Health England, 61 Colindale Avenue, London, NW9 5EQ
[2]University of East Anglia, Norwich, NR4 7TJ
[3]Animal Laboratories and Plant Health Agency, Woodham Lane, Surrey, KT15 3NB
[4]Division of Infection and Immunity, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, UK, EH25 9RG.
[5]Scottish *E. coli* O157/VTEC Reference Laboratory, Department of Laboratory Medicine, Royal Infirmary of Edinburgh, 51 Little France Crescent, Edinburgh EH16 4SA.
[6]Centre for Immunity, Infection and Evolution, Kings Buildings, University of Edinburgh, Edinburgh, UK, EH9 3FL.
[7] Future Farming Systems, R&D Division, SRUC, Drummondhill, Stratherrick Rd., Inverness, Scotland, UK, IV2 4JZ

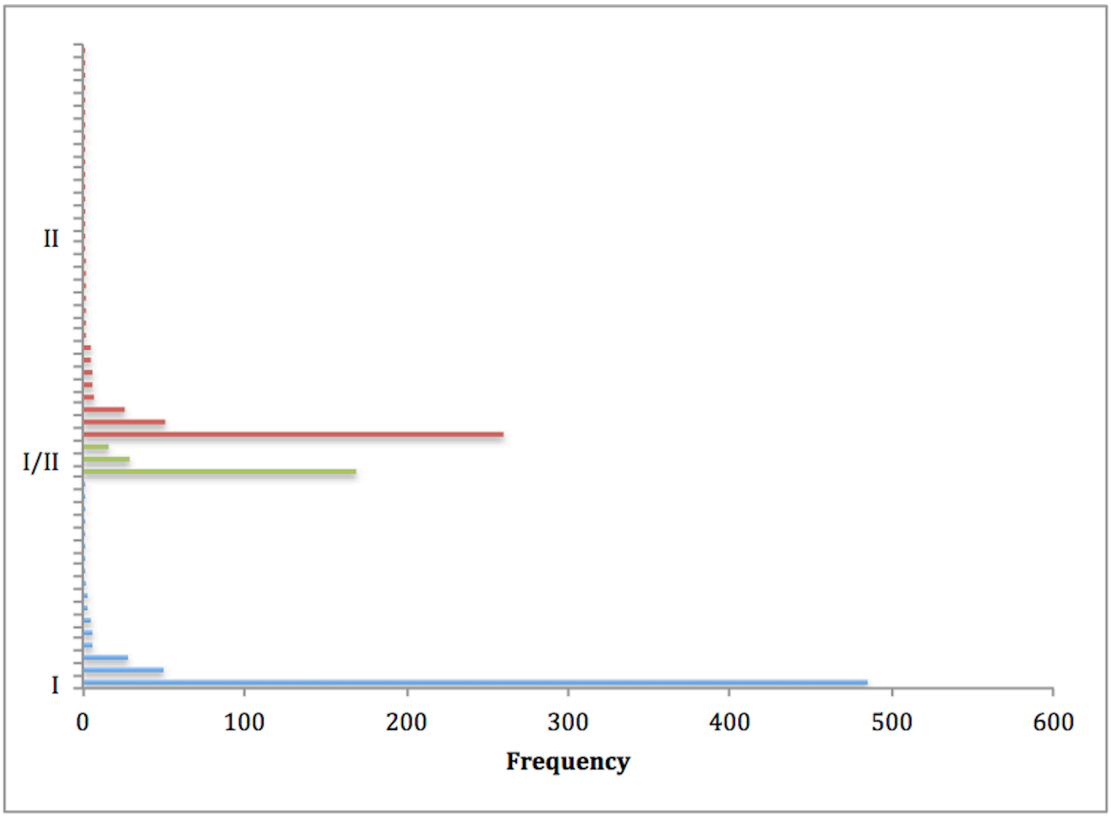
***Corresponding author – tim.dallman@phe.gov.uk**



Supplementary Figure 1a.
Maximum likelihood phylogeny of 584 STEC O157:H7 Δ25 SNP representatives depicting the three lineages. The outer circle is coloured by LSPA-6 type.



Supplementary Figure 1b.
Maximum likelihood phylogeny of 584 STEC O157:H7 Δ25 SNP representatives depicting the three lineages. The outer circle is coloured by Manning Clade type.



Supplementary Figure 2.
Bar chart showing the number of isolates in each Δ250 SNP cluster.

Lineage	Intra Cluster Recombination	Donor to Lineage I	Donor to Lineage II	Donor to Lineage I/II
I (828)	*19.3%*	*56.0%*	*20.3%*	*4.4%*
I/II (384)	*29.7%*	*15.1%*	*13.5%*	*41.7%*
II (1088)	*18.6%*	*16.5%*	*59.9%*	*5.0%*

Supplementary Table 1.
Table showing the direction of recombination for each of the three lineages of STEC O157:H7. The number in brackets represents the total number of donor segments per lineage.

Lineage II-a	80
stx 2c	**65**
yehV-sbcA	53
OTHER	12
stx negative	**6**
negative	5
yehV	1
stx 2a/2c	**4**
yecE-yehV-sbcA	2
wrbA-yehV-sbcA	1
Z2577-yecE-yehV-sbcA	1
stx 1a/2c	**3**
yehV-sbcA-argW	2
yehV-sbcA	1
stx 2a	**2**
yecE-yehV-sbcA	2

Lineage II-b	40
stx 2c	**20**
yehV-sbcA-argW	15
yehV-sbcA	4
yecE-yehV-sbcA-argW	1
stx negative	**8**
negative	8
stx 2a	**5**
yehV-sbcA-argW	4
yehV-sbcA	1
stx 1a/2c	**4**
sbcA-argW	2
yecE-sbcA-argW	1
yehV-sbcA-argW	1
stx 2a/2c	**3**
yehV-sbcA-argW	3

Lineage II-c	250
stx 1a/2c	**234**
yehV-sbcA	156
yehV-sbcA-argW	61
OTHER	17
stx 1a/2a/2c	**6**
yehV-sbcA	4
yecE-yehV-sbcA	2
stx 1a	**5**
yehV	5
stx 2a/2c	**3**
yehV-sbcA	3
stx 2a	**1**
yehV-sbcA-argW	1
stx 2c	**1**
yehV-sbcA	1

Supplementary Table 2:

The proportion of stx sub-type and occupied stx-associated bacteriophage insertion site (SBI) for each sub-lineage of lineage II. Those SBI's that less than 10% of the total were grouped into an 'other' category.

Lineage I/II	167
stx 2a/2c	**109**
yehV-sbcA-argW	88
OTHER	21
stx 2a	**55**
yehV-argW	44
yehV-sbcA-argW	6
OTHER	5
stx 2c	**3**
yecE-yehV-sbcA	2
yehV-argW	1

Supplementary Table 3:

The proportion of stx sub-type and occupied stx-associated bacteriophage insertion site (SBI) for lineage I/II. Those SBI's that less than 10% of the total were grouped into an 'other' category.

Lineage I-a	41
stx 1a/2a	**26**
wrbA-yehV	23
OTHER	3
stx 2a/2c	**6**
yehV-sbcA	6
stx 2c	**5**
yehV-sbcA	4
yecE-yehV-sbcA	1
stx 2a	**2**
wrbA-yehV	1
yecE-wrbA	1
stx 1a	**1**
yehV	1
stx negative	**1**
negative	1

Lineage I-b	30
stx 2c	**26**
yecE-yehV-sbcA	15
yecE-wrbA-yehV-sbcA	6
OTHER	5
stx 2a/2c	**1**
yecE-yehV-sbcA-argW	1
stx negative	**3**
negative	3

Lineage I-c	467
stx 2a/2c	**344**
yehV-sbcA-argW	265
OTHER	79
stx 2a	**88**
yehV-sbcA-argW	48
yecE-yehV	22
OTHER	18
stx 2c	**30**
yehV-sbcA	17
yehV-sbcA-argW	5
OTHER	8
stx 1a/2c	**3**
sbcA	1
yehV-sbcA	1
Z2577-yecE-wrbA-yehV-sbcA	1
stx negative	**2**
yehV-sbcA-argW	2

Supplementary Table 4:
The proportion of stx sub-type and occupied stx-associated bacteriophage insertion site (SBI) for lineage I. Those SBI's that less than 10% of the total were grouped into an 'other' category.

Supplementary Table 5:
All 1129 genomes analysed in this study summarised in terms of Lineage, SNP cluster, SBI, *stx* type, Manning Clade and LSPA-6 type.